

Split Sampling: Expectations, Normalisation and Rare Events

John R. Birge, Changgee Chang, and Nicholas G. Polson
Booth School of Business*

First Draft: August 2012
This Draft: October 2012

Abstract

In this paper we develop a methodology that we call split sampling methods which estimate expectations in high dimensions more precisely than other available methods. Split sampling methods are attractive for computing normalisation constants and estimating rare event probabilities. We implement our method using an auxiliary variable MCMC simulation with the expectation of interest expressed as an integrated set of rare event probabilities and derive our estimator from a Rao-Blackwellised estimate of a marginal auxiliary variable distribution. We illustrate our method with two applications. First, we compute a shortest path rare event probability and compare our method to estimation to a cross entropy approach. Then, we compute a normalisation constant of a high dimensional mixture of Gaussians and compare our estimate to one based on nested sampling. Finally, we discuss the relationship between our method and alternatives such as bridge sampling and the Wang-Landau algorithm. The methods developed here are available in the R package: **SplitSampling**.

Keywords: Rare Event, Normalisation, Cross Entropy, Slice Sampling, MCMC, Importance Sampling, Split Sampling, Serial Tempering, Annealing, Adaptive MCMC, Wang-Landau, Nested Sampling, Bridge Sampling.

*E-mail: john.birge@chicagobooth.edu, changgee@uchicago.edu, ngp@chicagobooth.edu. We would like to thank the participants at the conference in honor of Pietrio Muliere at Bocconi, September 13-15, 2012 for their helpful comments. The authors' work was supported by the University of Chicago Booth School of Business.

1 Introduction

In this paper we develop a methodology we refer to as split sampling methods to provide more precise estimates for expectations in high dimensions such as normalisation constants and rare event probabilities. We show that more precise estimators can be achieved by splitting the expectation of interest into a number of easier-to-estimate normalisation constants and then integrating those estimates to produce an estimate of the full expectation. To do this, we employ a family of splitting functions indexed by a random auxiliary variable and develop a weighting function with normalisation constants to generate the overall estimator. We allow for an adaptive MCMC approach to specify our weighting function. Other variance reduction techniques, such as control variates will provide further efficiency gains (see Dellaportas and Kontoyiannis, 2012, Mira et al, 2012).

There are two related approaches in the literature. One approach is nested sampling for estimation of normalisation constants which sequentially estimates the quantiles of the likelihood function under the prior to provide an estimator. Other normalisation methods include bridge and path sampling (Meng and Wong, 1996, Gelman and Meng, 1998), generalized versions of the Wang-Landau algorithm (Wang and Landau, 2001), and the TPA algorithm of Huber and Schott (2010). Serial tempering (Geyer, 2010) and linked importance sampling (Neal, 2005) provide ratios of normalisation constants for a discrete set of unnormalised densities. A second approach is cross entropy (Rubinstein and Glynn, 2009, Asmussen et al, 2012) which sequentially constructs an optimal variance-reducing importance function for calculating rare event probabilities. The product estimator (Diaconis and Holmes, 1994, Fishman, 1994) “splits” the rare event probability into a set of relatively large conditional probabilities which are easier to estimate.

We now introduce the notation to characterize the estimation problems and to develop our method. The central problem is to calculate an expectation of a positive functional of interest, which we denote as $L(\mathbf{x})$, under a k -dimensional probability distribution $\pi(\mathbf{x})$. We write this expectation as:

$$Z = \mathbb{E}_{\pi}(L(\mathbf{x})) = \int_{\mathcal{X}} L(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} .$$

The standard Monte Carlo estimate $\hat{Z} = (1/N) \sum_{i=1}^N L(\mathbf{x}^{(i)})$ where $\mathbf{x}^{(i)}$ is drawn from $\pi(\mathbf{x})$, possibly via MCMC, is too inaccurate. We introduce an auxiliary variable, m , a family of splitting functions, $L_m(\mathbf{x})$, and their normalisation constants, $Z_m = \int_{\mathcal{X}} L_m(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}$. For

rare events $L_m(\mathbf{x}) = \mathbb{I}(L(\mathbf{x}) > m)$; for normalisation constants $L_m(\mathbf{x}) = \exp(-mH(\mathbf{x}))$ and for counting problems $L_m(\mathbf{x}) = \#(\mathbf{x} : L(\mathbf{x}) > m)$. In general, the task at hand is NP-hard. We wish to develop a fully polynomial randomised approximation scheme.

We will focus on the case where the splitting functions $L_m(\mathbf{x})$ are specified by level sets in the likelihood, namely $L_m(\mathbf{x}) = \mathbb{I}(L(\mathbf{x}) > m)$. This occurs naturally in the rare event simulation literature and in nested sampling. The marginal distribution on m is defined via a weight function $\omega(m)$ with normalisation constants Z_m . The joint importance split distribution, $\pi_{SS}(\mathbf{x}, m)$, uses this marginal and the set of “tilted” posterior distributions. We assume that the researcher has a fast MCMC algorithm available for sampling the joint distribution.

The rest of the paper is outlined as follows. Section 2 details our split sampling methodology. We develop the key identity that “splits” the expectation of interest into an integrated set of rare event normalisation constants. MCMC then provides an estimator of the marginal distribution of the auxiliary variable, which in turn provides our overall estimator. We provide a number of guidelines for specifying our weight function. In the discrete case, we need to specify a “cooling schedule” for m , denoted by $\{m_t\}_{t=0}^T$, and adaptively estimate the weights to provide the greatest variance reduction. Section 3 describes the relationship with nested sampling methods. We show how to choose the weight function to mimic the sampling behavior of nested sampling. Section 4 applies our methodology to a shortest path rare event probability and to the calculation of a normalisation constant for a spike-and-slab mixture of Gaussians. We illustrate the efficiency gains of split sampling over crude Monte Carlo, the conditional probability estimator (Diaconis and Holmes, 1994, Fishman, 1994) and the cross entropy method. Finally, Section 5 concludes with an importance sampling view of nested and split sampling and directions for future research.

2 Split Sampling

Split sampling works as follows. Given a family of splitting functions $L_m(\mathbf{x})$ indexed by a random auxiliary variable m , we define the set of “tilted” distributions and corresponding normalisation constants by

$$\pi_m(\mathbf{x}) = \frac{L_m(\mathbf{x})\pi(\mathbf{x})}{Z_m} \text{ where } Z_m = \int_{\mathcal{X}} L_m(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} .$$

A case of particular interest occurs when $L_m(\mathbf{x}) = \mathbb{I}(L(\mathbf{x}) > m)$. In this case the tilted distribution $\pi_m(\mathbf{x}) \sim \pi(\mathbf{x}|L(\mathbf{x}) > m)$ corresponds to conditioning on level sets of $L(\mathbf{x})$. The Z_m 's correspond to “rare” event probabilities

$$Z_m = \int_{\mathcal{X}} L_m(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} = \int_{L(\mathbf{x}) > m} \pi(\mathbf{x})d\mathbf{x} = \mathbb{P}_\pi(L(\mathbf{x}) > m) .$$

We interpret $\pi(\mathbf{x})$ as a “prior” distribution, $L(\mathbf{x})$ as a likelihood, $\pi_M(x) = L(\mathbf{x})\pi(\mathbf{x})/Z$ as a “posterior” distribution. The key to our approach will be the specification and estimation of the marginal distribution of m .

The expectation of interest, Z , can be viewed as the integration of normalisation constants $\int_0^\infty Z_m dm$ via the key identity

$$Z = \int_{\mathcal{X}} L(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} = \int_0^\infty \left\{ \int_{m < L(\mathbf{x})} \pi(\mathbf{x})dm \right\} d\mathbf{x} = \int_0^\infty Z_m dm .$$

By this approach, we have “split” the computation of Z into a set of easier to compute normalisation constants Z_m .

2.1 Mixture Importance Splitting

Split sampling will simultaneously estimate \hat{Z}_m and $\hat{Z} = \int_0^\infty \hat{Z}_m dm$. First, we need to specify an appropriate mixture importance sampling distribution. Given weights, $\omega(m)$, we specify an importance splitting density on m by

$$\pi_{SS}(m) = \frac{\omega(m)Z_m}{\int_0^\infty \omega(m)Z_m dm} .$$

The joint distribution of $\pi_{SS}(\mathbf{x}, m)$ has conditional posterior, $\pi_{SS}(\mathbf{x}|m) \equiv \pi_m(\mathbf{x})$, and is

$$\pi_{SS}(\mathbf{x}, m) = \pi_{SS}(\mathbf{x}|m) \cdot \frac{\omega(m)Z_m}{\int_0^\infty \omega(m)Z_m dm} .$$

When $L_m(\mathbf{x}) = \mathbb{I}(L(\mathbf{x}) > m)$, the discrete joint distribution is given by

$$\pi_{SS}(\mathbf{x}, m_t) = \pi_{SS}(m_t)\pi_{SS}(\mathbf{x}|m_t) = \frac{\mathbb{I}(L(\mathbf{x}) > m_t)\pi(\mathbf{x})}{Z_{m_t}} \cdot \frac{\omega_t Z_{m_t}}{\sum_{t=0}^T \omega_t Z_{m_t}} .$$

The conditional distribution for $m = m_t$ given \mathbf{x} is given by

$$\pi(m_t|\mathbf{x}) = \frac{\omega_t \mathbb{I}_{\{L(\mathbf{x}) > m_t\}}}{\sum_{s=0}^T \omega_s \mathbb{I}_{\{L(\mathbf{x}) > m_s\}}}, \quad 0 \leq t \leq T.$$

The key feature of split sampling is that MCMC draws are available as is an efficient Rao-Blackwellised estimate of the marginal $\hat{\pi}_{SS}(m)$ without the knowledge of the Z_m 's. From the marginal, we will estimate \hat{Z}_m and $\hat{Z} = \int_0^\infty \hat{Z}_m dm$.

Samples from $(\mathbf{x}, m)^{(i)} \sim \pi_{SS}(\mathbf{x}, m)$ are available with Gibbs sampling which iterates the conditionals $\pi(\mathbf{x}|m) \sim \pi_m(\mathbf{x})$ and $\pi(m|\mathbf{x})$. When $L_m(\mathbf{x}) = \mathbb{I}(L(\mathbf{x}) > m)$, this reduces to running a Markov chain with density $\pi(\mathbf{x})$ conditioned to the set $\{\mathbf{x} : L(\mathbf{x}) > m\}$, namely

$$\pi(\mathbf{x}|m) \sim \pi(\mathbf{x}|L(\mathbf{x}) > m) .$$

We can sample $\pi(x_j|\mathbf{x}_{-j}, m)$ in a component-wise fashion as well. Additional auxiliary variables can be added to improve the speed of convergence (Polson, 1996).

Roberts (2010) observes that MCMC rather than Bayesian inference will achieve the largest efficiency gains in rare event probabilities (Glasserman et al, 1999, Glynn et al, 2010) and in counting problems where the resultant chains can be hard to sample exactly.

The complete conditional distribution for $m|\mathbf{x}$ does not depend on Z_m and is given by

$$\pi(m|\mathbf{x}) \propto \frac{L_m(\mathbf{x})\pi(\mathbf{x})}{Z_m} \omega(m) Z_m = \frac{\omega(m)L_m(\mathbf{x})}{\int_0^\infty \omega(m)L_m(\mathbf{x})dm} .$$

Given a discrete “cooling schedule”, denoted by $\{m_0, \dots, m_T\}$, we set $\omega(m) = \sum_{t=1}^T \omega_t \delta_{m_t}(m)$, where δ is a Dirac delta function. To achieve significant variance reduction gains, we will construct an adaptive MCMC method to determine an “optimal” schedule and weights. For rare events, we only require $Z(\gamma) = \mathbb{P}(L(\mathbf{x}) > \gamma)$ and we set $\omega(m) = 0, m > \gamma$.

2.2 The Estimator

We now derive estimators for Z_m and Z . By construction, the joint mixture importance splitting distribution is

$$\pi_{SS}(\mathbf{x}, m) = \frac{\omega(m)Z_m}{\int_0^\infty \omega(s)Z_s ds} \frac{L_m(\mathbf{x})\pi(\mathbf{x})}{Z_m} .$$

Given samples $\mathbf{x}^{(i)} \sim \pi_{SS}(\mathbf{x})$, we can exploit a Rao-Blackwellised estimator for the marginal via the identity $\pi_{SS}(m)$ which leads to the estimator

$$\begin{aligned}\hat{\pi}_{SS}(m) &= \mathbb{E}(\pi_{SS}(\mathbf{x}, m)) = \frac{1}{N} \sum_{i=1}^N \pi_{SS}(m | \mathbf{x}^{(i)}) \\ &= \frac{1}{N} \sum_{i=1}^N \frac{\omega(m) L_m(\mathbf{x}^{(i)})}{\int_0^\infty \omega(m) L_m(\mathbf{x}^{(i)}) dm} .\end{aligned}$$

Given our marginal density estimator, $\hat{\pi}_{SS}(m)$, we can estimate Z_m via

$$\frac{\hat{Z}_m}{Z_0} = \frac{\omega(0)}{\omega(m)} \frac{\hat{\pi}_{SS}(m)}{\hat{\pi}_{SS}(0)} .$$

As $Z_0 = 1$, it is natural to pick $\omega(0) = 1$. Our new estimate of \hat{Z}_m is then

$$\hat{Z}_m = \omega(m)^{-1} \cdot \frac{\hat{\pi}_{SS}(m)}{\hat{\pi}_{SS}(0)} .$$

Convergence of $\hat{Z}_m \rightarrow Z_m$ is straightforward. By the ergodic theorem, $\hat{\pi}_{SS}(m) \rightarrow \pi_{SS}(m) \forall m$. Therefore $\hat{\pi}_{SS}(m)/\hat{\pi}_{SS}(0) \rightarrow \pi_{SS}(m)/\pi_{SS}(0)$. The correction factor $\hat{\pi}_{SS}(m)/\hat{\pi}_{SS}(0)$ needs to be estimated as accurately as possible. To do this we will use an adaptive choice of the weight function, $\omega(m)$ and use convergence results from the adaptive MCMC literature.

When $L_m(\mathbf{x}) = \mathbb{I}(L(\mathbf{x}) > m)$, the splitting density is

$$\pi(\mathbf{x}, m) = \pi(m) \pi(\mathbf{x} | M = m) = \frac{\omega(m) Z_m}{\int_0^\infty \omega(s) Z(s) ds} \frac{\mathbb{I}_{\{L(\mathbf{x}) > m\}} \pi(\mathbf{x})}{Z_m} \quad 0 < m < \infty .$$

The posterior conditional of the auxiliary index m given \mathbf{x} is

$$\pi(m | \mathbf{x}) = \omega(m) \mathbb{I}(L(\mathbf{x}) > m) / \Omega(L(\mathbf{x})) \quad \text{where} \quad \Omega(m) = \int_0^m \omega(s) ds .$$

The marginal density estimator of m is

$$\begin{aligned}\hat{\pi}_{SS}(m) &= \frac{1}{N} \sum_{i=1}^N \pi(m | \mathbf{x}^{(i)}) = \frac{1}{N} \sum_{i=1}^N \frac{\omega(m) \mathbb{I}\{L(\mathbf{x}^{(i)}) > m\}}{\Omega(L(\mathbf{x}^{(i)}))} , \quad 0 < m < \infty \\ &= \phi_N(m) \omega(m) \quad \text{where} \quad \phi_N(m) = \frac{1}{N} \sum_{i=1}^N \frac{\mathbb{I}\{m < L(\mathbf{x}^{(i)})\}}{\Omega(L(\mathbf{x}^{(i)}))} .\end{aligned}$$

This is a re-weighted version of the initial weights $\omega(m)$.

The marginal density of \mathbf{x} is given by

$$\pi_{SS}(\mathbf{x}) = \frac{\Omega(L(\mathbf{x}))\pi(\mathbf{x})}{\int_0^\infty \omega(s)Z(s)ds}.$$

This provides a new estimator \hat{Z}_m , where $\mathbf{x}^{(i)} \sim \pi_{SS}(\mathbf{x})$, and

$$\hat{Z}_m = \frac{\omega(0)\hat{\pi}(m)}{\omega(m)\hat{\pi}(0)} = \frac{\sum_{\mathbf{x}^{(i)}: L(\mathbf{x}^{(i)}) > m} \Omega^{-1}(L(\mathbf{x}^{(i)}))}{\sum_{i=1}^N \Omega^{-1}(L(\mathbf{x}^{(i)}))}.$$

To find $\hat{Z} = \int_0^\infty \hat{Z}_m dm$, we use the summation-by-parts counterpart to Fubini, namely

$$\int_0^\infty \sum_{\mathbf{x}^{(i)}: L(\mathbf{x}^{(i)}) > m} \Omega^{-1}(L(\mathbf{x}^{(i)})) dm = \sum_{i=1}^N \Omega^{-1}(L(\mathbf{x}^{(i)})) L(\mathbf{x}^{(i)}).$$

Therefore, we have

$$\hat{Z} = \sum_{i=1}^N \frac{\Omega^{-1}(L(\mathbf{x}^{(i)}))}{\sum_{i=1}^N \Omega^{-1}(L(\mathbf{x}^{(i)}))} L(\mathbf{x}^{(i)}).$$

We now describe our split sampling algorithm.

Algorithm: Split Sampling

- Draw samples $(\mathbf{x}, m)^{(i)} \sim \pi_{SS}(\mathbf{x}, m)$ by iterating $\pi_{SS}(\mathbf{x}|m)$ and $\pi_{SS}(m|\mathbf{x})$
- Estimate the marginal distribution $\hat{\pi}_{SS}(m)$ via

$$\hat{\pi}_{SS}(m) = \frac{1}{N} \sum_{i=1}^N \frac{\omega(m)L_m(\mathbf{x}^{(i)})}{\int_0^M \omega(m)L_m(\mathbf{x}^{(i)})dm}.$$

- The new estimate of the individual normalisation constants, Z_m , is

$$\hat{Z}_m = \omega^{-1}(m) \cdot \frac{\hat{\pi}_{SS}(m)}{\hat{\pi}_{SS}(0)}.$$

- Compute a new estimate, \hat{Z} , via

$$\hat{Z} = \sum_{i=1}^N \frac{\Omega^{-1}(L(\mathbf{x}^{(i)}))}{\sum_{i=1}^N \Omega^{-1}(L(\mathbf{x}^{(i)}))} L(\mathbf{x}^{(i)}).$$

In the discrete case with a given grid, $\omega(m) = \sum_{t=0}^T \omega_t \delta_{m_t}(m)$. The marginal probabilities $\pi_t \equiv \pi(m_t)$ are estimated by Rao-Blackwellization as

$$\hat{\pi}_t = \frac{1}{N} \sum_{i=1}^N \pi(m_t | \mathbf{x}^{(i)}) = \frac{1}{N} \sum_{i=1}^N \frac{\omega_t \mathbb{I}_{\{L(\mathbf{x}^{(i)}) > m_t\}}}{\sum_{s=0}^T \omega_s \mathbb{I}_{\{L(\mathbf{x}^{(i)}) > m_s\}}}, \quad 0 \leq t \leq T.$$

With $Z_0 = 1$, the estimator is $\hat{Z}_t = \omega_0 \hat{\pi}_t / \omega_t \hat{\pi}_0$ for $0 \leq t \leq T$.

There is an equivalence with the product estimator where $Z = Z_0 \prod_{t=1}^T Z_{m_t} / Z_{m_{t-1}}$ over a discrete grid m_t of m -values. Variance reduction is achieved by splitting Z into pieces $Z_{m_t} / Z_{m_{t-1}}$ of larger magnitude which are relatively easier to estimate.

With $\mathbf{x}_t^{(i)} \sim \pi_{m_{t-1}}(\mathbf{x})$, we estimate

$$\frac{Z_{m_t}}{Z_{m_{t-1}}} = \int_{\mathcal{X}} \frac{L_{m_t}(\mathbf{x})}{L_{m_{t-1}}(\mathbf{x})} \pi_{m_{t-1}}(\mathbf{x}) d\mathbf{x} \quad \text{with} \quad \widehat{\frac{Z_{m_t}}{Z_{m_{t-1}}}} = \frac{1}{N} \sum_{i=1}^N \frac{L_{m_t}(\mathbf{x}_t^{(i)})}{L_{m_{t-1}}(\mathbf{x}_t^{(i)})}.$$

Given N independent samples from T tilted distributions, we have

$$\hat{Z} = \prod_{t=1}^T \frac{1}{N} \sum_{i=1}^N \frac{L_{m_t}(\mathbf{x}_t^{(i)})}{L_{m_{t-1}}(\mathbf{x}_t^{(i)})}.$$

The product estimator, as well as the cross-entropy estimator, relies on a set of independent samples drawn in a sequential fashion. Split sampling, on the other hand, uses a fast MCMC and ergodic averaging to provide an estimate \hat{Z} . The Monte Carlo standard error $\text{var}(\hat{\pi}_{SS}(m) / \hat{\pi}_{SS}(0))$ can be determined from the output of the chain.

Controlling the Monte Carlo error is straightforward due to independent samples with relative mean squared error, e.g. Garvels et al (2002), given by

$$\hat{Z} = \prod_{t=1}^T \widehat{\frac{Z_{m_t}}{Z_{m_{t-1}}}} \quad \text{and} \quad \frac{\text{Var}(\hat{Z})}{\mathbb{E}(\hat{Z})^2} = \prod_{t=1}^T \left(\frac{\hat{\sigma}_{m_t}^2}{\hat{\mu}_{m_t}^2} + 1 \right) - 1$$

with mean $\mathbb{E}(\hat{Z}_{m_t} / \hat{Z}_{m_{t-1}}) = \hat{\mu}_{m_t}$ and variance $\hat{\sigma}_{m_t}^2$. Picking m_t such that the coefficients of variation are all the same leads to an estimator with reduced variance.

Huber and Schott (2010) discusses the theoretical advantages of constructing a “well-balanced” cooling schedule. Using Chebyshev-Hoeffding bounds, we obtain a running time of $O((\ln(1/Z))^2)$ steps. Stefankovic, Vempola and Vigoda (2009) provide theoretical arguments that can reduce the running times to $O(\ln(1/Z))$ steps. The empirical test of our mixture IS algorithm is whether there exists a fast MCMC algorithm for sampling $\pi(\mathbf{x}, m)$ that reduces the $O(NT)$ effort in sequential methods.

2.3 Choice of $\omega(m)$

The previous subsection assumed that $\omega(m)$ is fixed. To improve the efficiency of our algorithm, we can use an adaptive MCMC approach to specifying the weight function.

A common initialisation is to set $\omega(m) \equiv 1, \forall m$. Then, $\Omega(L(\mathbf{x})) = \int_0^{L(\mathbf{x})} \omega_s ds = L(\mathbf{x})$. This leads to an estimate of the marginal, $\mu_N(m) \equiv \hat{\pi}_{SS}(m)$, given by the measure

$$\mu_N(m) = \left\{ \frac{1}{N} \sum_{i=1}^N \frac{\mathbb{I}\{m < L(\mathbf{x}^{(i)})\}}{\Omega(L(\mathbf{x}^{(i)}))} \right\} \omega(m) .$$

The density estimate will be zero for $m > L(\max_i \mathbf{x}^{(i)})$ by construction. We also have a set of estimates $\hat{Z}_m^{-1} = \sum_{m < L(\mathbf{x}^{(i)})} L^{-1}(\mathbf{x}^{(i)}) / \sum_{i=1}^N L^{-1}(\mathbf{x}^{(i)})$ where $\mathbf{x}^{(i)} \sim \pi_{SS}(\mathbf{x})$.

Given an initial run of the algorithm, we can re-proportion the prior weight function to regions we have not visited frequently enough. To accomplish this, let $\phi(m)$ be a desired target distribution for $\pi_{SS}(m)$, for example a uniform measure. Then re-balance the weights inversely proportional to the visitation probabilities and set the new weights $\omega^*(m)$ by

$$\frac{\omega^*(m)}{\omega(m)} = \frac{\omega(m)}{\mu_N(m)} = \frac{\mathbb{I}\{m < L(\mathbf{x}^{(i)})\}}{\Omega(L(\mathbf{x}^{(i)}))} .$$

This will only adjust our weights in the region where $m < \max_i L(\mathbf{x}^{(i)})$. As the algorithm proceeds we will sample regions of higher likelihood values and further adaptative our weight function.

Other choices for weights are available. For example, in many normalisation problems Z_m will be exponential in m due to a Laplace approximation argument. This suggests taking an exponential weighting $\omega(m) = \kappa e^{\kappa m}$ for some $\kappa > 0$. In this case, we have

$$\Omega(m) = \int_0^{m \wedge M} \omega(s) ds = e^{\kappa(m \wedge M)} - 1 .$$

The marginal distribution is

$$\pi_{SS}(\mathbf{x}) = \frac{(e^{\kappa(L(\mathbf{x}) \wedge M)} - 1) \pi(\mathbf{x})}{\int_0^\infty \kappa e^{\kappa s} Z_s ds} .$$

We can also specify $\omega(m)$ to deal with the possibility that the chain might not have visited all states by setting a threshold ω_{max} which corresponds to the maximum allowable increase in the log-prior weights. This leads to a re-balancing rule

$$\frac{\omega^*(m)}{\omega(m)} = \min \left\{ \frac{\max_m \mu_N(m)}{\mu_N(m)}, e^{\omega_{max}} \right\} ,$$

where we have also re-normalised the value of the largest state to one.

When L_{max} is available, we set $M = L_{max}$ and $\omega(m) = 0$ for $m > M$. To initialise $\omega(m)$, we use the harmonic mean $\hat{Z}_{L_{max}}^{-1}$ for $\omega(M)$ and an exponential interpolation for $\omega(m)$. Drawing $\mathbf{x}^{(i)} \sim \pi_M(\mathbf{x}) = L(\mathbf{x})\pi(\mathbf{x})/Z$, we have

$$Z_M^{-1} = E_{\pi_M}(L_M^{-1}(\mathbf{x})) \quad \text{to estimate} \quad \hat{\omega}(M) = \frac{1}{N} \sum_{i=1}^N L_M^{-1}(\mathbf{x}^{(i)}) .$$

The harmonic mean estimator (Raftery et al, 2007) is known to have poor Monte Carlo error variance properties (Polson, 2006, Wolpert and Schmidler, 2012) although we are estimating $\omega(m)$ and not its inverse.

We can extend this insight to a fully adaptive update rule for $\omega_N(m)$, similar to stochastic approximation schemes. Define a sequence of decreasing positive step sizes γ_n with $\sum_{n=1}^{\infty} \gamma_n^{-1} = \infty, \sum_{n=1}^{\infty} \gamma_n^{-2} < \infty$. A practical recommendation is $\gamma_n = Cn^{-\alpha}$ where $\alpha \in [0.6, 0.7]$, see e.g. Sato and Ishii (2000). Another approach is to wait until a “flat histogram” (FH) condition holds:

$$\max_{m \in \{m_t\}} \left| \mu_N(m) - \phi(m) \right| < c .$$

for a pre-specified tolerance threshold, c . The measure $\mu_N(m) = (1/N) \sum_{i=1}^N \#(m^{(i)} = m)$ tracks our current estimate of the marginal auxiliary variable distribution. The Rao-Blackwellised estimate $\hat{\pi}_{SS}(m)$ further reduces variance.

The empirical measure can be used to update $\omega_N(m)$ as the chain progresses. Let κ_N denote the points at which γ_{κ_N} will be decreased according to its schedule. Then an update rule whci guarantees convergence is to set

$$\log \omega_{\kappa_N}(m) \leftarrow \log \omega_{\kappa_N-1}(m) + \gamma_{\kappa_N} (\mu_{\kappa_N}(m) - \phi(m)) .$$

Jacob and Ryder (2012) show that if γ_N is only updated on a sequence of values κ_N which correspond to times that a “flat-histogram” criterion is satisfied, then convergence ensues and the FH criteria is achieved in finite time. After updating $\omega_{\kappa_N}(m)$, we re-set the counting measure $\mu_{\kappa_N}(m)$ and continue. Other adaptive MCMC convergence methods are available in Atchade and Liu (2010), Liang et al (2007), and Zhou and Wong (2008). Bornn et al (2012) provides a parallelisable algorithm for further efficiency gains. Peskun (1973) provides theoretical results on optimal MCMC chains to minimise the variance of MCMC functionals.

One desirable Monte Carlo property for an estimator is a bounded coefficient of variation. For simple functions, $L(\mathbf{x}) = \mathbf{x}$ and $\max_i \mathbf{x}_i$, mixture importance functions achieve such a goal, see Iyengar (1991) and Adler et al (2008). Madras and Piccioni (1999, section 4) hint at the efficiency properties of dynamically selected mixture importance blankets. Gramacy et al (2010) propose the use of importance tempering. Johansen et al (2006) use logit annealing implemented via a sequential particle filtering algorithm.

2.3.1 Discrete Cooling Schedule

We suggest a simple, sequential, empirical approach to selecting a “cooling schedule” in our approach. Specifically, set $m_0 = 0$, then given m_{t-1} we sample $\mathbf{x}^{(i)} \sim \pi_{m_{t-1}}(\mathbf{x}) \sim \pi(\mathbf{x}|L(\mathbf{x}) > m_{t-1})$. We order the realisations of the criteria function $L(\mathbf{x}^{(i)})$ and set m_t equal to the $(1 - \rho)$ -quantile of the $L(\mathbf{x}^{(i)})$ samples. This provides a sequential approach to solving

$$\rho = \mathbb{P}(L(\mathbf{x}) > m_t | L(\mathbf{x}) > m_{t-1}) = \mathbb{P}(L(\mathbf{x}) > m_t) / \mathbb{P}(L(\mathbf{x}) > m_{t-1}) = Z_{m_t} / Z_{m_{t-1}}.$$

A number of authors have proposed “optimal” choices of ρ , which implicitly defines a cooling schedule, m_t , for $0 \leq t \leq T$. L’Ecuyer et al (2006) and Amrein and Kunsch (2011) propose $\rho = e^{-2}$ and 0.2, respectively. Huber and Schott (2010) define a well-balanced schedule as one that satisfies $e^{-1} < \rho < 2e^{-1}$. They show that such a choice leads to fast algorithms. The difficulty is in finding the right order of magnitude of M and the associated schedule m_t that ensures that each slice $Z_{m_t}/Z_{m_{t-1}}$ is not exponentially small. For rare events, we sample until $m_{t+1} > M$ and then set $m_T = M$. Our initial estimate $\hat{Z} = \rho^{-T}$ and our weights are $\omega(m) = \rho^m$.

In hard cases, such as the multimodal mixture of Gaussians, the normalising constants $Z(m)$ are not exponential in m . In such cases we initialize the weights by a piecewise exponential obtained by interpolating any point $m \in (m_{t-1}, m_t)$ by $\Omega(m) = \Omega_{t-1} \exp(\kappa_t(m - m_{t-1}))$ where $\kappa_t = \log(\Omega_t/\Omega_{t-1})/(m_t - m_{t-1})$. For $m > m_T$, we use $\Omega(m) = \Omega_T$.

2.4 Discussion

2.4.1 Importance Sampling Interpretation

The estimator \hat{Z}_m can be viewed as an importance sampling estimator where we average $\Omega^{-1}(L(\mathbf{x}^{(i)}))$ over the splitting set $L(\mathbf{x}^{(i)}) > m$ with $\mathbf{x}^{(i)} \sim \pi_{SS}(\mathbf{x})$. Similarly we can express \hat{Z} as an importance sampling estimator as follows. Given $\omega(s)$ and $\Omega(m) = \int_0^m \omega(s)ds$, consider importance sampling with a blanket proportional to $\Omega(L(\mathbf{x}))$. Then

$$Z = \int_{\mathcal{X}} L(\mathbf{x})\pi(\mathbf{x}) = \int_{\mathcal{X}} \Omega^{-1}(L(\mathbf{x}))L(\mathbf{x}) \left\{ \int_0^{L(\mathbf{x})} \omega(s)ds \right\} \pi(\mathbf{x})d\mathbf{x}. \quad (1)$$

Split sampling uses a proposal distribution proportional to $\Omega(L(\mathbf{x}))\pi(\mathbf{x})$. This can be viewed as the marginal from a weighted slice distribution

$$\pi_{SS}(\mathbf{x}, m) = \frac{\omega(s)\pi(\mathbf{x})}{\int_0^\infty \omega_s Z_s ds} \text{ on } 0 \leq s \leq L(\mathbf{x}).$$

The estimator of Z , from (1), can be viewed as the ratio of normalisation constants

$$\frac{Z}{\int_0^\infty \omega_s Z_s ds} = \int_{\mathcal{X}} \Omega^{-1}(\mathbf{x})\pi_{SS}(\mathbf{x}) \approx \frac{1}{N} \sum_{i=1}^M \Omega^{-1}(\mathbf{x}^{(i)}) \text{ where } \mathbf{x}^{(i)} \sim \pi_{SS}(\mathbf{x}).$$

To estimate the normalisation constant, $\int_0^\infty \omega_s Z_s ds$, we use the harmonic mean

$$1/\frac{1}{N} \sum_{i=1}^N \Omega^{-1}(\mathbf{x}^{(i)}) \text{ where } \mathbf{x}^{(i)} \sim \pi_{SS}(\mathbf{x}).$$

Therefore, our importance sampling estimate becomes

$$\hat{Z} = \sum_{i=1}^N \frac{\Omega^{-1}(L(\mathbf{x}^{(i)}))}{\sum_{i=1}^N \Omega^{-1}(L(\mathbf{x}^{(i)}))} L(\mathbf{x}^{(i)}).$$

We also observe that the marginal ordinate $\pi_{SS}(0) = 1/\int_0^\infty \omega_s Z_s ds$ with $\omega_0 = 1$. With this interpretation, if the MCMC algorithm spends significant time in the zero state of m , then we will be able to provide an accurate Monte Carlo estimate of $\pi_{SS}(0)$ and hence of Z .

2.4.2 Adaptive Mixture Interpretation

Our methodology can be viewed as an adaptive mixture importance sampler. Umbrella sampling (Torrie and Valleau, 1997) can be seen as a precursor to many of the current

advanced MC strategies such as the Wang-Landau algorithm and its generalisations for sampling high dimensional multimodal distributions. These algorithms exploit an auxiliary variable and by their adaptive nature improve estimates continuously as the simulation advances. The main difference is how each algorithm traverses low and high energy states. The Wang-Landau algorithm aims to achieve a uniform distribution on the auxiliary variable, thus spending more time in low energy states than high states as opposed to multicanonical sampling (Berg and Neuhaus, 1992), $1/k$ -ensemble sampling (Hesselbo and Stinchcombe, 1995) or simulated tempering (Geyer and Thompson, 1995).

The Wang Landau algorithm was originally designed for discrete probability distributions. Here the state space is split into M disjoint bins; so, $\mathcal{X} = \cup_{i=1}^K \mathcal{X}_i$. When $\pi(\mathbf{x}) = \exp(-mH(\mathbf{x})) / Z_m$, the algorithm provides estimates of the level energy sets $\#\{\mathbf{x} : H(\mathbf{x}) = u\}$. Liang (2005) provides a generalisation to continuous distributions. As in equi-energy sampling (Kuo et al, 2006), we define $\mathcal{X}_i = \{\mathbf{x} : u_{i-1} < H(\mathbf{x}) < u_i\}$ and the algorithm provides estimates of $g_i = \int_{\mathcal{X}_i} \pi(\mathbf{x}) d\mathbf{x}$. Both algorithms can be viewed as a discrete mixture importance splitter with $H_m(x) = \mathbb{I}_{\mathcal{X}_m}(\mathbf{x})$.

The equilibrium distribution $\pi_{WL}(x)$ spends equal time in each bin, thus disproportionately sampling low energy states. The target density is specified by weights $\psi_t = \int_{\mathcal{X}_t} \pi(\mathbf{x}) d\mathbf{x}$ and is given by $\pi_\psi(\mathbf{x}) \propto \pi(\mathbf{x}) / \psi_{J(\mathbf{x})}$ where $J(\mathbf{x})$ is the index such that $\mathbf{x} \in \mathcal{X}_{J(\mathbf{x})}$. As the objective is to estimate (ψ_1, \dots, ψ_K) , the Wang-Landau approach uses an adaptive MCMC algorithm based on the sequence of densities $\pi_{\theta_t}(\mathbf{x}) \propto \pi(\mathbf{x}) / \theta_t(J(\mathbf{x}))$ where θ_t is adaptively updated using a stochastic approximation scheme or a “flat-histogram” criterion. Fort et al (2012) discuss convergence properties of the Wang-Landau algorithm.

2.4.3 Bridge Sampling Interpretation

Much of the intuition of split sampling methods occurs when $m \in \{0, 1\}$ and we have a two component mixture of prior and posterior. Given $\omega > 0$, define

$$\pi_{SS}(\mathbf{x}) = \frac{1}{1 + \omega Z} \pi(\mathbf{x}) + \frac{\omega Z}{1 + \omega Z} \frac{L(\mathbf{x}) \pi(\mathbf{x})}{Z}$$

We will estimate the relative probability ωZ of being in each component. This is related to the Savage-Dickey approach to calculating Bayes factors.

To see this, we write the conditional distribution of the mixture indicator is

$$\pi(m|\mathbf{x}) \sim \text{Ber}(p(\mathbf{x})) \text{ where } p(\mathbf{x}) = \frac{\pi(m=1|\mathbf{x})}{\pi(m=0|\mathbf{x})} = \omega Z \frac{L(\mathbf{x}) \pi(\mathbf{x}) / Z}{\pi(\mathbf{x})} = \omega L(\mathbf{x}).$$

As in slice sampling, we view $\pi_{SS}(\mathbf{x})$ as a marginal from the joint distribution

$$\pi_{SS}(\mathbf{x}, m) = \frac{1}{1 + \omega Z} \pi(\mathbf{x}) + \frac{\omega Z}{1 + \omega Z} \frac{\mathbb{I}(L(\mathbf{x}) > m) \pi(\mathbf{x})}{Z}.$$

3 Comparison with Nested Sampling

We now describe nested sampling (NS, Skilling, 2006). Here we use the π density to calculate the ordered x -quantiles of the likelihood. Then if we assume that $z(\mathbf{x})$ has a well defined inverse $\mathbf{x}(z)$ we have

$$z(\mathbf{x}) = Z_{L(\mathbf{x})} = \int_{L(\mathbf{x}') > L(\mathbf{x})} \pi(\mathbf{x}') d\mathbf{x}'$$

Under this change of variables, the equivalent identity approximates the expectation of interest, Z , via simple quadrature

$$Z = \int_0^1 L(\mathbf{x}(z)) dz \approx \sum_{t=1}^N L_t(x_t - x_{t-1}).$$

Nested sampling sequentially determines L_t by sampling $\pi(\mathbf{x}|L(\mathbf{x}) > L_{t-1})$. Brewer et al (2011) propose a diffuse nested sampling approach to determine the levels L_t . Both nested and diffuse nested sampling are product estimator approaches. The quantiles $0 = L_0 < \dots < L_t < \dots$ are chosen so that each level L_t occupies $\rho = e^{-1}$ times as much prior mass as the previous level L_{t-1} . Diffuse nested sampling achieves this by sequentially sampling from a mixture importance sampler $\sum_{j=1}^{t-1} w_j \mathbb{I}(L(\mathbf{x}) > L_{j-1}) \pi(\mathbf{x})$ where the weights are exponential $w_j \propto e^{\kappa(j-t)}$ for some κ . MCMC methods are used to traverse this mixture distribution with a random walk step for the index j that steps up or down a level with equal probability. A new level is added using the $(1 - e^{-1})$ -quantile of the likelihood draws. Using diffuse nested sampling allows some chance of the samples' escaping to lowered constrained levels and to explore the space more freely. One caveat is that a large contribution can come from values of $\mathbf{x}(z)$ near the origin and we have to find many levels T to obtain an accurate approximation.

Murray et al (2006) provide single and multiple sample versions of nested sampling algorithms. If $L_{\max} = \sup_{\mathbf{x}} L(\mathbf{x})$ is known, we sample as follows: set $X = 1, N = 1, Z = 0$.

1. Generate $\mathbf{x}^{(1)} \sim \pi(\mathbf{x})$ and set $L_0 = 0, L_1 = L(\mathbf{x}^{(1)})$.

2. If $L_{\max}X < \epsilon Z$, then set $Z = Z + (X/N) \sum_{j=1}^N L_{i+j-1}$ and stop.

3. Repeat while $L_i X/N > \delta$:

(a) Generate $\mathbf{x}^{(i+N)} \sim \pi(\mathbf{x}) \mathbb{I}(L(\mathbf{x}) > L_{i-1})$ and set $L_i = L(\mathbf{x}^{(i+N)})$.

(b) Set $N = N + 1$ and sort L_i 's.

4. Set $Z = Z + L_i X/N$ and $N = N - 1$, $X = (1 - 1/N)X$.

If L_{\max} is not known, replace step 2 with:

2(a) If $L_{i+N-1}X < \epsilon Z$, then set $Z = Z + (X/N) \sum_{j=1}^N L_{i+j-1}$.

We now choose $\omega(m)$ to match the sampling properties of nested sampling. The main difference between split and nested sampling is that in split sampling we specify a weight function $\omega(m)$ for $0 < m < \infty$ and sample from the full mixture distribution, rather than employing a sequential approach for grid selection which requires a termination rule. Another difference is that split sampling estimator does not need to know the ordered L_t 's.

3.1 Matching Split and Nested Sampling

We take the sampling distribution for nested sampling from Skilling (2006). The expected number of samples less than m is $-N \log Z(m)$. If $N = 1$, we have $Z(L_i)/Z(L_{i-1})$ are independent standard uniforms since

$$-\log Z(L_k) = -\sum_{i=1}^k \log \frac{Z(L_i)}{Z(L_{i-1})} \quad \text{for } k \geq 1.$$

The distribution of the number of samples less than m is the number of arrivals before $-\log Z(m)$ of a Poisson process with rate 1. This generalises to arbitrary N .

The sampling distribution of the nested sampling for finite n is hard to calculate, but we can observe limiting results. As $n \rightarrow \infty$, if $N/n \rightarrow \lambda$, then we have

$$Z_{NS}(m) = \mathbb{P}^{NS}(L(\mathbf{x}) > m) = 1 + \lim_{n, N \rightarrow \infty} \frac{N}{n} \log Z(m) = 1 + \lambda \log Z(m).$$

Split sampling has marginal density of \mathbf{x} given by

$$\pi_{SS}(\mathbf{x}) = \frac{\Omega(L(\mathbf{x}))\pi(\mathbf{x})}{\int_0^\infty \omega(s)Z(s)ds} \quad \text{with } \Omega(m) = \int_0^m \omega(s)ds.$$

The tail distribution function $Z_{SS}(m) = \pi_{SS}(L(\mathbf{x}) > m)$ is then

$$\begin{aligned} Z_{SS}(m) &= \int_0^\infty \int_{\mathcal{X}} \mathbb{I}_{\{L(\mathbf{x}) > m\}} \frac{\omega(s)Z(s)}{\int_0^\infty \omega(s)Z_s ds} \frac{\mathbb{I}_{\{L(\mathbf{x}) > s\}}\pi(\mathbf{x})}{Z_s} d\mathbf{x} ds \\ &= \frac{\int_0^\infty \omega(s)Z(m \vee s) ds}{\int_0^\infty \omega(s)Z_s ds} = \frac{Z_m \Omega(m) + \int_m^\infty \omega(s)Z_s ds}{\int_0^\infty \omega(s)Z_s ds}. \end{aligned}$$

We now find the importance splitting density that matches the nested sampling distributional properties in the sense that $Z_{NS}(m) = Z_{SS}(m)$. Since

$$Z'_{NS}(m) = \frac{\lambda Z'_m}{Z_m} \quad \text{and} \quad Z'_{SS}(m) = \frac{Z'_m \Omega(m)}{\int_0^\infty \omega(s)Z_s ds},$$

where $Z'_{NS}(m) = \partial Z_{NS}(m) / \partial Z_m$. We therefore set $\Omega(m) = Z^{-1}(m)$.

As we wish to traverse the full likelihood surface, we also introduce a condition that lets us monitor the number of visits, N_{level} , to the current top level of the likelihood before we construct a new level. Specifically, we run

1. Set $T = 0$, $m_0 = 0$, $\Omega_0 = 1$, and $Z_0 = 1$.
2. While $T < T_{max}$, set $T = T + 1$
 - (a) Simulate $\pi_{SS}(\mathbf{x})$ with $\{m_t\}_{0 \leq t < T}$ and $\{\Omega_t\}_{0 \leq t < T}$ until we have N_{level} visits to level $T - 1$.
 - (b) Choose the $(1 - \rho)$ -quantile of likelihoods of level $T - 1$ as m_T .
 - (c) Set $\hat{Z}_T = \rho^{-T}$.
 - (d) Set $\Omega_T = \hat{Z}_T^{-1}$.

Under the condition $\Omega(m) = Z(m)^{-1}$, the chain will visit each level roughly uniformly. However, it may take a long time to reach the top level, and the uncertainty in Ω may act like a hurdle for visiting upper levels. With these concerns, it is desirable to favor upper levels by replacing step (d) with

- (d1) Set $\Omega_T = e^{\Lambda T} Z_T^{-1}$.

We call Λ the boosting factor as Λ increases the preference for the upper levels. This reduces the search time and ensures the time complexity to be $O(T)$. To further expedite this procedure, we may put more weight on the top level T by substituting (d) with the step:

(d2) Set $\Omega_{T-1} = e^{\Lambda(T-1)} Z_{T-1}^{-1}$ and $\Omega_T = \beta \frac{e^{\Lambda T} - 1}{e^{\Lambda} - 1} Z_T^{-1}$.

For example, if $\beta = 1$, the chain spends half of the time on the top level and the other half backtracking the other levels.

Once we identify all levels, our split sampling algorithm runs:

1. Set $i = 0$ and $\nu_t = \nu_{init} \hat{Z}_t$ for each $t = 0, 1, \dots, T$.
2. While $i \leq n$, set $i = i + 1$.
 - (a) Draws $\mathbf{x}^{(i)}$ using MH under weights $\Omega(m)$, and set $L_i = L(\mathbf{x}^{(i)})$.
 - (b) Obtain $M^{(i)} = \Omega^{-1}(U_i)$ where $U_i \sim U(0, \Omega(L_i))$ with $\Omega(0) = 1$.
 - (c) For each t with $m_t < L_i$, update $\nu_t = \nu_t + \Omega(L_i)^{-1}$.
 - (d) Update $\hat{Z}_t = \nu_t / \nu_0$ and set $\Omega_t = \hat{Z}_t^{-1}$.

4 Applications

4.1 Rare Event Shortest Path

Calculating rare event probabilities is a common goal of many problems. Rubinstein and Kroese (2004) consider the total length of the shortest path on a weighted graph with random weights $\mathbf{x} = (x_1, \dots, x_5)$. Each weight x_j follows an independent exponential distribution with scale parameter u_j with joint distribution given by

$$\pi(\mathbf{x}|u) = \left(\prod_{j=1}^5 \frac{1}{u_j} \right) \exp \left(- \sum_{j=1}^5 \frac{x_j}{u_j} \right) \text{ where } u = (0.25, 0.4, 0.1, 0.3, 0.2) .$$

The goal is to estimate the probability of the rare event corresponding to the length of the shortest path

$$Z(\gamma) = \mathbb{P}(S(\mathbf{x}) > \gamma) \text{ where } S(\mathbf{x}) = \min(x_1 + x_4, x_1 + x_3 + x_5, x_2 + x_3 + x_4, x_2 + x_5)$$

We will consider three cases: $\gamma = 2, 3$, and 4 where the true rare event probabilities are

$$Z(2) = 1.34 \times 10^{-5}, \quad Z(3) = 2.06 \times 10^{-8}, \quad \text{and} \quad Z(4) = 3.10 \times 10^{-11}.$$

These can be estimated by the split sampler (SS) with $L(\mathbf{x}) = S(\mathbf{x})$ and level breakpoints $\{0 = m_0, m_1, \dots, m_T = \gamma\}$. $\hat{Z}_T \equiv \hat{Z}(m_T)$ is the estimator.

We implement three other competing estimators. First, the crude Monte Carlo (CMC) estimator simulates $\mathbf{x}^{(i)} \sim \pi(\mathbf{x}|\mathbf{u})$ and estimates the rare event probabilities by

$$\hat{Z}(\gamma) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{\{S(\mathbf{x}^{(i)}) > \gamma\}}.$$

Second, the conditional probability product (CPP) estimator $\hat{Z}(\gamma)$ calculates the $(1 - \rho)$ -quantile m_{t+1} of N_0 samples of $S(\mathbf{x}^{(t,i)})$ under $\mathbf{x}^{(t,i)} \sim \pi_t(\mathbf{x}) \propto \mathbb{I}_{\{S(\mathbf{x}) > m_t\}} \pi(\mathbf{x})$ for all $t = 0, \dots, T-1$ with $m_0 = 0$, $m_{T-1} < \gamma$, and $m_T \geq \gamma$. This estimator is defined as:

$$\hat{Z}(\gamma) = \left(\prod_{t=1}^{T-1} \frac{\hat{Z}(m_t)}{\hat{Z}(m_{t-1})} \right) \frac{\hat{Z}(\gamma)}{\hat{Z}(m_{T-1})} = \rho^{T-1} \frac{1}{N_0} \sum_{i=1}^{N_0} \mathbb{I}_{\{S(\mathbf{x}^{(T-1,i)}) > \gamma\}}.$$

To find $\mathbf{x}^{(t,i)}$ we need to sample $\pi(\mathbf{x}|S(\mathbf{x}) > m_t)$. We use Gibbs sampling with complete conditionals $\pi(x_i|\mathbf{x}_{(-i)}, S(\mathbf{x}) > m)$ given by truncated exponential distributions. By the lack of memory property, we have

$$\pi(x_1|\mathbf{x}_{(-1)}, S(\mathbf{x}) > m) = \max(0, m - x_4, m - x_3 - x_5) + x_1^* \text{ where } x_1^* \sim \text{Exp}(u_1).$$

The other conditionals $\pi(x_i|\mathbf{x}_{(-i)}, S(\mathbf{x}) > m)$ follow in a similar manner.

Third, the *cross-entropy* (CE) estimator (de Boer et al, 2005) calculates an “optimal” importance blanket, $\pi(\mathbf{x}|\hat{\mathbf{v}}_T)$, parameterised by $\hat{\mathbf{v}}_T$. Then it draws N_1 samples of $\mathbf{x}^{(i)} \sim \pi(\mathbf{x}|\hat{\mathbf{v}}_T)$ and estimates the shortest path probability

$$\hat{Z}(\gamma) = \frac{1}{N_1} \sum_{i=1}^{N_1} \mathbb{I}_{\{S(\mathbf{x}^{(i)}) > \gamma\}} w(\mathbf{x}^{(i)}; \mathbf{u}, \hat{\mathbf{v}}_T), \text{ where } w(\mathbf{x}^{(i)}; \mathbf{u}, \hat{\mathbf{v}}_T) = \frac{\pi(\mathbf{x}^{(i)}|\mathbf{u})}{\pi(\mathbf{x}^{(i)}|\hat{\mathbf{v}}_T)}.$$

The sequential algorithm for finding $\hat{\mathbf{v}}_T$ is similar in spirit to the product estimator approach: set $\hat{\mathbf{v}}_0 = \mathbf{u}$ and $t = 1$. Choose ρ ; typically $\rho = 0.1$. Then perform

1. Draw N samples of $\mathbf{x}^{(i)} \sim \pi(\mathbf{x}|\hat{\mathbf{v}}_{t-1})$. Let $\hat{\gamma}_t$ be the $(1 - \rho)$ quantile of $S(\mathbf{x}^{(i)})$.

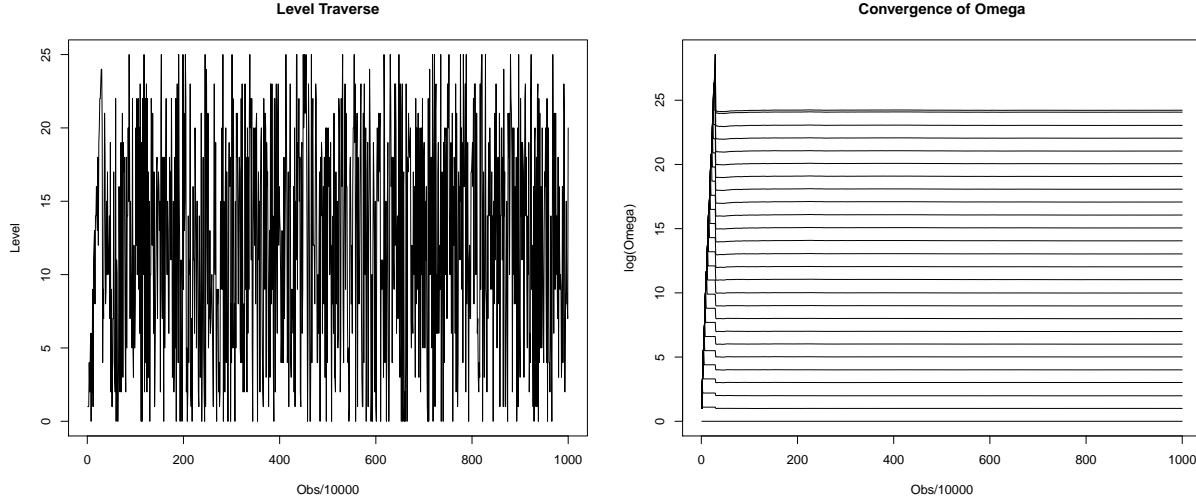
If $\hat{\gamma}_t > \gamma$, set $\hat{\gamma}_t = \gamma$.

2. Update $\hat{\mathbf{v}}_{t-1}$ via cross-entropy minimisation;

$$\hat{\mathbf{v}}_t = \frac{\sum_{i=1}^N \mathbb{I}_{\{S(\mathbf{x}^{(i)}) > \hat{\gamma}_t\}} w(\mathbf{x}^{(i)}; \mathbf{u}, \hat{\mathbf{v}}_{t-1}) \mathbf{x}^{(i)}}{\sum_{i=1}^N \mathbb{I}_{\{S(\mathbf{x}^{(i)}) > \hat{\gamma}_t\}} w(\mathbf{x}^{(i)}; \mathbf{u}, \hat{\mathbf{v}}_{t-1})}.$$

Table 1: Rare event probabilities simulation

N	$\gamma = 2$				$\gamma = 3$			$\gamma = 4$		
	CMC	CE	CPP	SS	CE	CPP	SS	CE	CPP	SS
10^5	0.807	0.040	0.044	0.055	0.066	0.076	0.091	0.113	0.098	0.133
10^6	0.275	0.011	0.015	0.015	0.017	0.025	0.026	0.028	0.033	0.036
10^7	0.086	0.003	0.004	0.005	0.005	0.007	0.007	0.008	0.009	0.011

Figure 1: Rare event $\gamma = 4$

3. If $\hat{\gamma}_t = \gamma$, set $T = t$ and exit. Otherwise, set $t = t + 1$ and go to step 1.

Table 1 provides the simulation results. Each scenario was run 100 times and *relative* RMS of each estimator was recorded. The CMC estimator was only recorded for $\gamma = 2$ as the other events are too rare to even have a single count. The total sample size, N , was 10^5 , 10^6 , or 10^7 . The tuning parameters for CE were $\rho = 0.1$, $N_0 = 1,000$ for $\gamma = 2$ and $N_0 = 10,000$ for $\gamma = 3, 4$, and $N_1 = N - TN_0$. For CPP, we used $\rho = e^{-1}$ and $N_0 = N/T$ where T is the number of required steps. For split sampling (SS), we set $\rho = e^{-1}$, $N_{level} = 10,000$ and $\nu_{init} = 10,000$ and $\Lambda = 0.1$. Cross-entropy (CE) method finds an efficient importance sampling function as does split sampling. Further gains from split sampling are expected in higher dimensional problems where finding \hat{v}_t at each stage in CE can be cumbersome in general.

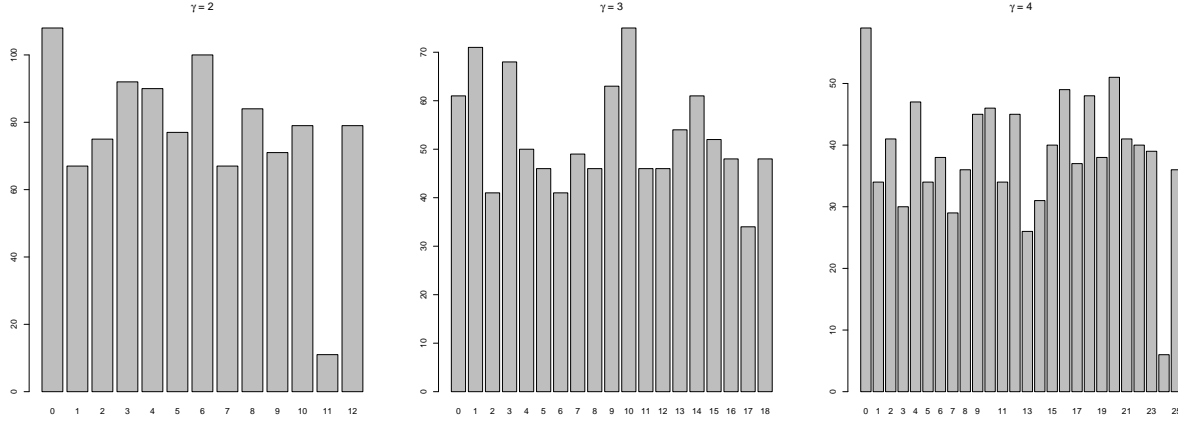


Figure 2: Histogram of Levels

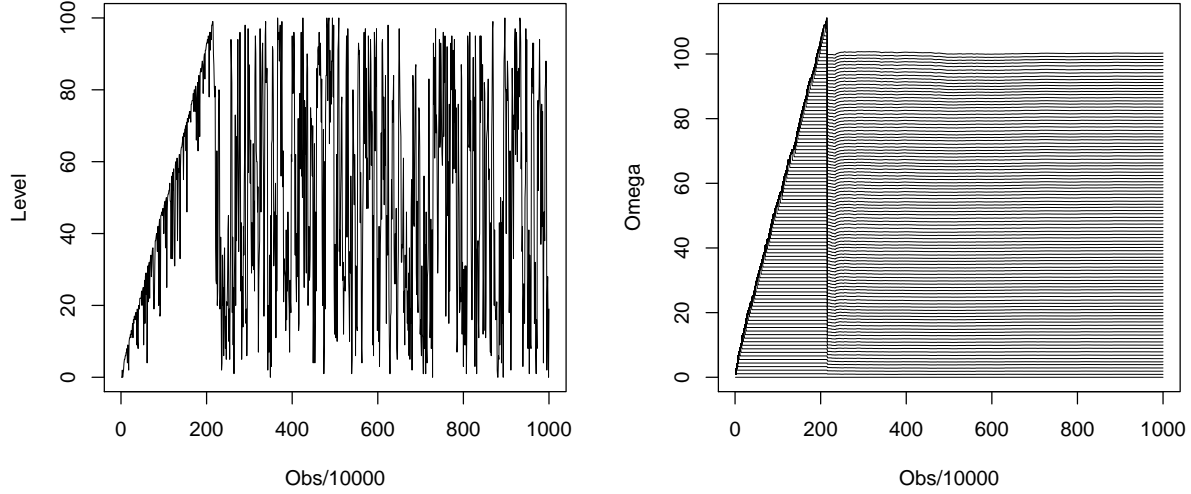


Figure 3: Centered Gaussians. Learning $\omega(m)$

4.2 Normalisation of a Mixture of Gaussians

As an illustration of the advantages of using split sampling we consider a centered and de-centered mixture of Gaussians. We follow the nested and diffuse nested sampling literature (Skilling, 2008, Brewer et al, 2011) and suppose that $\mathbf{x} = (x_1, \dots, x_C)$ where $C = 20$.

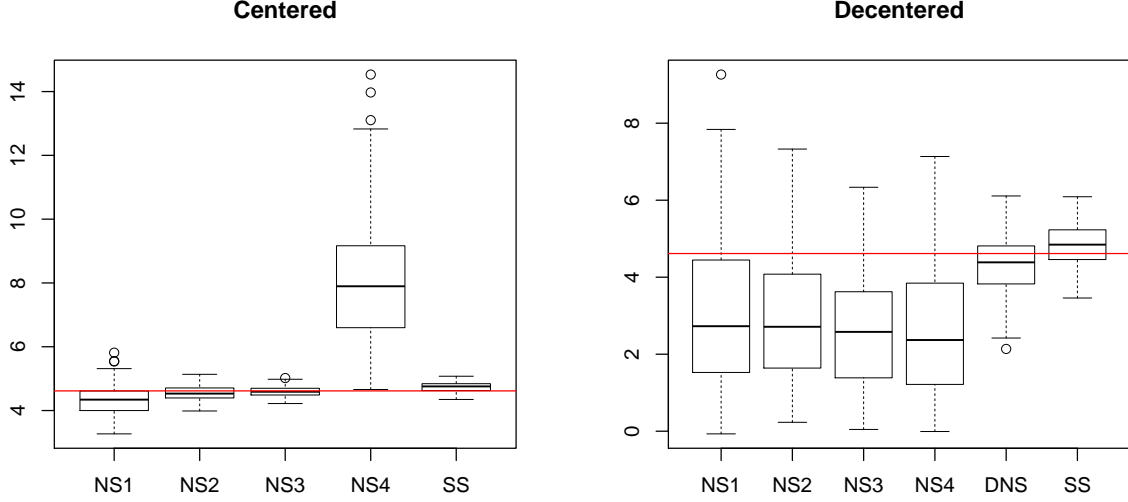


Figure 4: NS vs DNS vs SS

The centered likelihood is given by the classic Gaussian “spike-and-slab” of width 0.01 and “plateau” of width 0.1, namely

$$L_C(\mathbf{x}) = 100 \prod_{i=1}^{20} \frac{1}{\sqrt{2\pi}u} e^{-\frac{x_i^2}{2u^2}} + \prod_{i=1}^{20} \frac{1}{\sqrt{2\pi}v} e^{-\frac{x_i^2}{2v^2}},$$

The prior $\pi(\mathbf{x})$ is uniform. In the de-centered multimodal mixture we take

$$L_{DC}(\mathbf{x}) = 100 \prod_{i=1}^{20} \frac{1}{\sqrt{2\pi}u} e^{-\frac{(\mathbf{x}_i - 0.031)^2}{2u^2}} + \prod_{i=1}^{20} \frac{1}{\sqrt{2\pi}v} e^{-\frac{\mathbf{x}_i^2}{2v^2}}.$$

The goal is to calculate the so-called evidence, $Z = \int L(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} = 101$.

For the centered case, nested sampling vastly outperforms annealing and provides Monte Carlo bounds given by $\log Z \approx \log(101) \pm \sqrt{H/N}$ with $H = \int_{\mathcal{X}} \log(\pi_L(\mathbf{x})/\pi(\mathbf{x})) \pi_L(\mathbf{x})d\mathbf{x} = 63.2$. For the de-centered case, diffuse nested sampling is preferable.

For the de-centered case, the rare event normalising constants $Z(m)$ are severely L -shaped. When the likelihood is multimodal, the chain needs to be able to go down one likelihood to traverse another mode. The mixture split sampling importance blanket is designed to achieve this goal. One caveat, however, is that a poor set of initial probabilities

Table 2: SS vs NS, $u = 0.01$, $v = 0.1$, centered at origin, rms of $\log(\hat{Z})$ with true value $\log(Z) = 4.615$. The number of MCMC steps is reported per each NS step.

Algorithm	Parameters	RMS
NS1	300 particles, 333 MCMC steps	0.557
NS2	1000 particles, 100 MCMC steps	0.260
NS3	3000 particles, 33 MCMC steps	0.174
NS4	10000 particles, 10 MCMC steps	3.647
SS	$\rho = e^{-1}$, $T_{max} = 100$, $\nu = 5000$, $\Lambda = 10$	0.207

and weights, $\Omega(m) = Z(m)^{-1}$, leads to an algorithm that results in very low probability estimates for the levels on the spike. The “flat-histogram” criteria described below can take a long time to fix this problem. We ensure an adequate number of samples at the last level by placing more weight on the final level by multiplying ω_T by T .

We use the split sampling algorithm as described in (2.3.1). If the initial \hat{Z}_t are not as accurate as needed to guarantee good mixing of our MCMC iterations, we dynamically refine Ω_t as in step (d). The beauty of this scheme is that this update makes the chain self-balanced. When Ω_t is larger than it should be, or Z_t is smaller, the chain visits level t more often. Thus increasing Z_t and decreases Ω_t , which helps Z_t converge more quickly to the true value.

From a practical perspective, it is critical to have nonzero initial values on ν_t . If we start with $\nu_t = 0$, the early \hat{Z}_t and Ω_t are unstable, and the whole procedure can become abortive. Although not very accurate, $\hat{Z}_t = \rho^{-t}$ is a reasonable initial estimate and we sample long enough so the chain stabilizes. ν_{init} represents the degree of dependence on those initial values.

The final estimator is $\hat{Z} = \nu_t/\nu_0$ and

$$Z = \int_0^\infty Z(m)dm = \sum_{t=1}^T \int_{m_{t-1}}^{m_t} Z_{t-1} \exp(-\kappa_t(m - m_{t-1}))dm = \sum_{t=1}^T \frac{(Z_t - Z_{t-1})(m_t - m_{t-1})}{\log Z_t - \log Z_{t-1}}.$$

At first sight, the time complexity appears to be $O(nT)$ since steps (c)-(e) involve $O(T)$ operations. However, if the m_t values are chosen so that Z_t are exponentially decreasing, the work can be done in $O(n \log T)$ time. The expected number of comparisons in step (c) is the reciprocal of the decreasing rate of Z_t . The updates needed at steps (d) are only for the

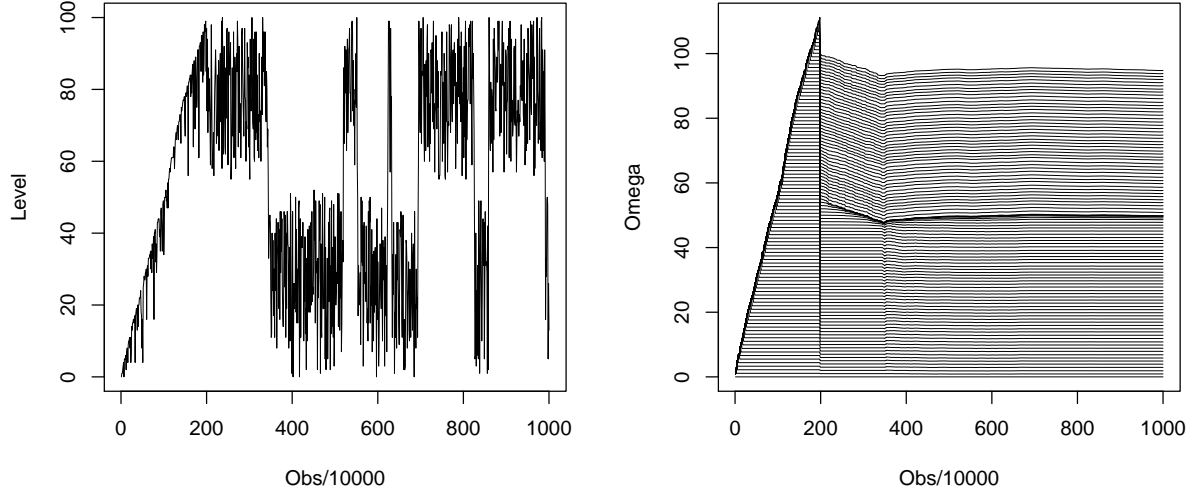


Figure 5: Decentered Gaussians. Learning $\omega(m)$

last several t 's since the increment $\Omega(L_i)^{-1}$ becomes negligible very quickly relative to ν_t as t decreases. We use a random walk MH proposal distribution given by $x_j^* = x_j + N(0, \sigma^2)$ where the density of the step size $f(\sigma) \propto 1/\sigma$ on $[10^{-4.5}, 10^{1.5}]$ for random chosen index j .

Table 3: SS vs (D)NS, $u = 0.01$, $v = 0.1$, the spike centered at $(0.031, \dots, 0.031)$, rms of $\log(\hat{Z})$ with $\log(Z) = 4.615$. The number of MCMC steps is per each NS step. Diffuse nested sampling (DNS, Brewer et al, 2011).

Algorithm	Parameters	RMS
NS1	300 particles, 333 MCMC steps	2.467
NS2	1000 particles, 100 MCMC steps	2.338
NS3	3000 particles, 33 MCMC steps	2.519
NS4	10000 particles, 10 MCMC steps	2.620
DNS	Diffuse Nested Sampling	0.763
SS	$\rho = e^{-1}$, $T_{max} = 100$, $\nu = 5000$, $\Lambda = 10$	0.591

Table 2 compares the performance of the nested sampling and the split sampling methods. For each run, $\log \hat{Z}$ were recorded and their root mean squares are reported. The

number of runs for each case is 500. For nested sampling, the sample size n is random and we instead fix N such that the average sample size becomes slightly greater than the target sample size n .

We also compare split sampling with nested and diffuse nested sampling in Table 3. Split sampling provides a RMS error of 0.591 versus 0.763 for diffuse nested sampling.

5 Discussion

von Neumann's original view of importance sampling was a method for variance reduction. By viewing the calculation of an expectation as a problem of normalising a posterior distribution, we can write

$$\pi_L(\mathbf{x}) = L(\mathbf{x})\pi(\mathbf{x})/Z \text{ where } Z = \int_{\mathcal{X}} L(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} .$$

Importance sampling uses a blanket $g(\mathbf{x})$ to compute

$$Z = \int_{\mathcal{X}} L(\mathbf{x}) \frac{\pi(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x} \approx \frac{1}{N} \sum_{i=1}^N L(\mathbf{x}^{(i)}) \frac{\pi(\mathbf{x}^{(i)})}{g(\mathbf{x}^{(i)})} \text{ where } \mathbf{x}^{(i)} \sim g(\mathbf{x}) .$$

Picking $g(\mathbf{x})$ to be the posterior distribution $L(\mathbf{x})$ leads to the estimator $\hat{Z} = Z$ with zero variance. While impractical, this suggests finding a class of importance blankets $g(\mathbf{x})$ that are adaptive and depend on $L(\mathbf{x})$ can exhibit good Monte Carlo properties.

Split sampling specifies a class of importance sampling blankets, indexed by $\omega(m)$, by

$$g_{\omega}(\mathbf{x}) = \frac{\left\{ \int_0^{L(\mathbf{x})} \omega_s ds \right\} \pi(\mathbf{x})}{\int_{\mathcal{X}} \Omega(\mathbf{x}) \pi(\mathbf{x}) d\mathbf{x}} \text{ where } \Omega(L(\mathbf{x})) = \int_0^{L(\mathbf{x})} \omega_s ds .$$

The advantage of the class of split sampling densities is that the resultant estimator of Z can be implemented via an auxiliary MCMC algorithm from a joint distribution $\pi_{SS}(\mathbf{x}, m)$ indexed by a random auxiliary variable m . Moreover, it allows an adaptive choice of $\omega_N(m)$ to reduce the Monte Carlo error of the resultant estimator. Convergence results rely on adaptive MCMC literature.

Split sampling illustrates the adaptive importance sampling nature of nested sampling and cross-entropy methods. There is also a clear relationship with slice sampling (Polson, 1996, Neal, 2003) as one can view the sampling of the posterior, $\pi_L(\mathbf{x})$, as the marginal

from the augmented distribution $\pi(\mathbf{x}, m) = \mathbb{I}(L(\mathbf{x}) > m) \pi(\mathbf{x})/Z$. The main difference is that split sampling runs a Markov chain that traverses the whole space defined by (\mathbf{x}, m) to find regions where $\omega(m)$ needs to be refined. Both CE and NS methods using a sequential sampling procedure as in the CPP estimator to split the quantity of interest into estimable pieces. Further research is required to tailor the specification of the weight function $\omega(m)$ to the problem at hand.

We leave open the question of an optimal choice of $L_m(\mathbf{x})$. Here we have focused on $L_m(\mathbf{x}) = \mathbb{I}(L(\mathbf{x}) > m)$, however, using logit-type functions might lead to faster converging MCMC algorithms. The key to the efficiency of split sampling is being able to construct a rapidly mixing MCMC algorithm to sample the mixture distribution $\pi_{SS}(\mathbf{x}, m)$. We aim to report on direct applications in Bayesian inference in future work. For example, Murray et al (2006) shows that nested sampling performs well for Markov random fields models and split sampling should have similar properties.

6 References

- Adler, R.J., J. Blanchet and J. Liu (2008). Efficient simulation for tail probabilities of Gaussian random fields. *Proceedings of the Winter Simulation Conference*, 328-336.
- Amrein, M. and H. Künsch (2011). A variant of importance sampling for rare event estimation: fixed number of successes. *ACM Transactions on Modeling and Computer Simulation*, 21(2), Article 13.
- Asmussen, S. and P. Dupuis, R. Rubenstein and H. Wang (2012). Importance Sampling for Rare Events. *Working Paper*.
- Atchadé, Y.F. and J.S. Liu (2010). The Wang-Landau algorithm in general state spaces: applications and convergence analysis. *Statistica Sinica*, 20, 209-233.
- Berg, B.A. and T. Neuhaus (1992). Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Physical Review Letters*, 68, 9-12.
- Bornn, L., P.E. Jacob, P. Del Moral and A. Doucet (2012). An adaptive interacting Wang-Landau algorithm for automatic density exploration. *Working Paper*.
- de Boer, P.T., D.P. Kroese and R.Y. Rubenstein (2005). A Tutorial on the cross-entropy method. *Annals of Operations Research*, 134, 19-67.

- Brewer, B.J, L.B. Pártay and G. Csányi (2011). Diffusive Nested Sampling. *Statistics and Computing*, 21(4), 649-656.
- Dellaportas, P. and I. Kontoyiannis (2012). Control variates for estimation based on reversible MCMC samplers. *Journal of Royal Statistical Society, B*, 74(1), 133-161.
- Diaconis, P. and S. Holmes (1994). Three examples of the MCMC method. *Discrete Probability and Algorithms*, 43-56.
- Fishman, G. (1994). Markov chain sampling and the Product Estimator. *Operations Research*, 42(6), 1137-1145.
- Fort, G., B. Jourdain, E. Kuhn, T. Leïèvre and G. Stoltz (2012). Convergence and efficiency of the Wang-Landau algorithm. *Working Paper*.
- Garvels, M.J.J., J.C.W. van Ommeren and D.P. Kroese (2002). On the importance function in splitting simulation. *European Transactions on Telecommunications*, 13(4), 363-371.
- Gelman, A. and X. Meng (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13, 163-185.
- Geyer, C.J. and E.A. Thompson (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of American Statistical Association*, 90, 909-920.
- Geyer, C.J (2012). Bayes factors via Serial Tempering. *Technical Report*.
- Glasserman, P., P. Heidelberger, P. Shahabuddin and T.Zajic (1999). Multi-Level Splitting for rare event probabilities. *Operations Research*, 47(4), 585-600.
- Glynn, P.W., A. Dolgin, R.Y. Rubinstein and R. Vaisman (2010). How to generate uniform samples on discrete sets using the splitting method. *Prob. Eng. Info. Sci.*, 24(3), 405-422.
- Gramacy, R., R. Samworth and R. King (2010). Importance Tempering. *Statistics and Computing*, 20(1), 1-7.
- Hesselbo, B. and R.B. Stinchcombe (1995). Monte Carlo Simulation and global optimisation without parameters. *Phys. Rev. Lett.*, 74, 2151-2155.
- Huber, M. and S. Schott (2010). Using TPA for Bayesian Inference. *Bayesian Statistics*, 9, 257-282.
- Iyengar, S. (1991). Importance Sampling for Tail Probabilities. *Technical Report 440*, Stanford University.

- Jacob, P.E. and R.J. Ryder (2012). The Wang-Landau algorithm reaches the Flat Histogram criterion in finite time. *Working paper*.
- Johansen, A.M., P. Del Moral and A. Doucet (2006). Sequential Monte Carlo Samplers for Rare Events. *Proceedings of the 6th International Workshop on Rare Event Simulation*, 256-267.
- Kuo, S.C., Q. Zhou and W.H. Wong (2006). Equi-energy sampler with applications in statistical inference and statistical mechanics. *Annals of Statistics*, 34(4), 1581-1619.
- L'Ecuyer, P., V. Demers and B. Tuffin (2006). Splitting for Rare event simulation. *Proceedings of the 2006 Winter Simulation Conference*, 137-148.
- Liang, F. (2005). Generalized Wang-Landau algorithm for Monte Carlo computation. *Journal of American Statistical Association*, 100, 1311-1337.
- Liang, F., C. Liu and R.J. Carroll (2007). Stochastic approximation in Monte Carlo computation. *Journal of American Statistical Association*, 102, 305-320.
- Madras, N. and M. Piccioni (1999). Importance Sampling for Families of Distributions. *Annals of Applied Probability*, 9(4), 1202-1225.
- Meng, X-L and W. Wong (1996). Simulating ratios of normalising constants via a simple identity: a theoretical exposition. *Statistica Sinica*, 6, 831-860.
- Mira, A., R. Solgi, and D. Imparato (2012). Zero Variance MCMC for Bayesian Estimators. *Statistics and Computing*.
- Murray, I., D.J.C. MacKay, Z. Ghahramani and J. Skilling (2006). Nested sampling for the Potts models. *Advances in NIPS*, 947-954.
- Neal, R.M. (2003). Slice Sampling. *Annals of Statistics*, 31(3), 705-767.
- Neal, R.M. (2005). Estimating ratios of normalising constants using linked importance sampling. *Technical Report No. 0511*, University of Toronto.
- Peskun, P.H. (1973). Optimum Monte Carlo sampling using Markov Chains. *Biometrika*, 60(3), 607-612.
- Polson, N.G. (1996). Convergence of Markov Chain Monte Carlo Algorithms. *Bayesian Statistics*, 5, 297-321.
- Polson, N.G. (2006). Comment on: "Estimating the integrated likelihood in posterior simulation using the harmonic mean equality". *Bayesian Statistics*, 8, 415-417.

- Raftery, A.E., M.A. Newton, J.M. Satagopan and P.N. Krivitsky (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean equality. *Bayesian Statistics*, 8, 371-417.
- Roberts, G.O. (2010). Comment on: “Using TPA for Bayesian Inference”. *Bayesian Statistics*, 9, 280-282.
- Rubinstein, R.Y. and P.W. Glynn (2009). How to deal with the curse of dimensionality of likelihood ratios in Monte Carlo simulation. *Stochastic Models*, 25, 547-568.
- Rubenstein, R.Y. and D. Kroese (2004). *The Cross Entropy Method*. Springer.
- Sato, M. and S. Ishii (2000). On-line EM algorithm for the normalized Gaussian network. *Neural Computation*, 12, 407-432.
- Skilling, J. (2006). Nested Sampling for General Bayesian Computation. *Bayesian Analysis*, 1(4), 833-860.
- Skilling, J. (2008). Nested Sampling for Bayesian computation. *Bayesian Statistics*, 8, 491-507.
- Štefankovič, D., S. Vempola and E. Vigoda (2009). Adaptive simulated annealing: a near-optimal connection between sampling and counting. *Journal of the ACM*, 56(3), Article No. 18.
- Torrie, G.M. and J.P. Valleau (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella Sampling. *Journal of Computational Physics*, 23(2), 187-199.
- Wang, F. and D.P. Landau (2001). Efficient multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett*, 86, 2050-2053.
- Wolpert, R.L. and S.C. Schmidler (2012). α -Stable limit laws for Harmonic Mean estimators of marginal likelihoods. *Statistica Sinica*, 22, 1233-1251.
- Zhou, Q. and W.H. Wong (2008). Reconstructing the energy landscape of a distribution from Monte Carlo samples. *Annals of Applied Statistics*, 2(4), 1307-1331.